# HIBA VAGY VÉTSÉG?
# ÖNBECSAPÁS ÉS ERKÖLCSI FELELŐSSÉG*

*BERNÁTH LÁSZLÓ – NYÁRFÁDI KRISZTIÁN*

A tanulmánynak két fő célja van. Az első részben bemutatja az önbecsapás *intencionális* és *nem-intencionális* magyarázatainak alapjait, amely magyarázati sémák jelenleg a legnépszerűbbek a kortárs analitikus filozófiában. A második részben az önbecsapás nem-intencionális elméleteinek egyik nehézsége kerül a fókuszba: hogyan lehet az önbecsapó felelős az önbecsapásért, amennyiben feltesszük, az önbecsapás során egyáltalán nem állt szándékában önmagát becsapni. Amellett fogunk érvelni, hogy az önbecsapás paradigmatikus eseteiben akkor is lehetséges az erkölcsi felelősség, ha az önbecsapás során önmagunk megtévesztése egyáltalán nem szándékos.

## 1. HAZUGSÁG ÉS ÖNBECSAPÁS

Képzeljük el a következő szituációt: Sam erőteljesen vonzódik egy csoporttársához, Sallyhez, akivel rendszeresen együtt szokott a vizsgákra készülni. Viszonyuk bizalmas, a lány sok, a magánéletét érintő kérdésbe beavatja őt. Sam leghőbb vágya, hogy Sally viszontszeresse őt. Úgy véli, a vonzalom kölcsönös. Egy megfelelőnek tűnő pillanatban randevúra is hívja a lányt, ám az visszautasítja közeledését. Kiegyensúlyozott párkapcsolatára hivatkozik, továbbá biztosítja Samet arról, hogy – bár nagyon hízelgő az ajánlata – ő bizony sajnálatos módon nem viszonozza a fiú gyengéd érzelmeit. Indítványozza továbbá, hogy maradjanak barátok. Sam mindezek ellenére továbbra is azt hiszi, hogy Sally szerelmes belé.

Vegyünk egy másik példát: egy anyát telefonon értesít a rendőrség arról, hogy fiát súlyos vádakkal előzetes letartóztatásba helyezték. Mikor bemegy a rendőrőrsre, a kirendelt ügyvéd szembesíti a fia ellen szóló tárgyi bizonyítékokkal, majd biztosítja, hogy – bár minden tőle telhetőt megtesz – a fiú néhány év szabadságvesztésre mindenképp számíthat. Az anya mindezek ellenére szentül meg van győződve fia ártatlanságáról.

Tegyük fel, hogy mind Sam, mind pedig az anya tévednek. Sally tényleg nem viszonozza Sam gyengéd érzelmeit, a fiú pedig tényleg bűnös. Tegyük fel továbbá, hogy sem Sam, sem pedig az anya nem szenvednek semmilyen

* Köszönettel tartozunk Kovács Gábornak, Makai Ádámnak, Tihanyi Katalinnak, Such Dávidnak, Tőzsér Jánosnak és különösen Szalai Juditnak kritikus észrevételeikért.

kognitív deficitben, amelyek lehetetlenné tennék számukra, hogy a rendelkezésre álló bizonyítékokat mérlegre téve igaz hiteket alakítsanak ki. Mégsem ezt teszik. Világos, hogy tévedéseikben szerepet játszanak a vágyaik. Ezeket az eseteket szokás mind a hétköznapokban, mind pedig a filozófiatörténetben önbecsapásként diagnosztizálni.

Ha a jelenség explicit filozófiatörténeti előfordulásait keressük, szinte bizonyosan a morálfilozófia területére tévedünk. Az önbecsapás Platóntól Kanton át egészen Butlerig erkölcsi hanyatlásunk biztos jele (Kantnál az önmagunkkal szembeni kötelességek megszegése, Butlernél „belső képmutatás"). Az utóbbi évtizedekben azonban az önbecsapás (és általában az irracionális jelenségek) szakirodalmában, elsősorban angolszász nyelvterületen, egy igen markáns elmozdulás figyelhető meg. A morálfilozófiai kérdéseket háttérbe szorították a transzcendentális, illetve a magyarázati kérdések („Vajon miképpen lehetséges egyáltalán az önbecsapás?" „Milyen folyamatok vezetnek hozzá?)". Mindazonáltal alapvetően továbbra is az „önmagunknak való hazugság" gondolata uralta a jelenség kapcsán folytatott filozófiai vitákat.

A hagyományos nézet konceptuális rokonságot lát a *személyközi megtévesztés* és az önbecsapás jelensége között. A személy becsapja B személyt akkor, ha A tudja (vagy legalábbis ténylegesen hiszi), hogy *nem-p*, és szánt szándékkal elhiteti B személlyel, hogy *p*. Ennek két feltétele:

1. Az egymásnak ellentmondó hitek kritériuma: a folyamat eredményeképpen A azt hiszi, hogy *nem-p*,[1] B pedig azt, hogy *p*. A megtévesztés paradigmatikus eseteiben lényegi elem, hogy a két személy végül egymásnak ellentmondó hitekkel rendelkezzen.

2. Szándékoltság: a megtévesztés per definitionem szándékolt cselekvés.

Hogy a személyközi megtévesztés modellje milyen mértékben szolgáltathat analógiát az önbecsapás jelenségének magyarázatakor, óriási viták tárgyául szolgál az angolszász filozófusok körében, legkésőbb a 60-as évektől egészen napjainkig. Az ezen modelltől való eltérés mértéke alapján szokás megkülönböztetni két alapvető magyarázati modellt: az *intencionális* és a *nem-intencionális* modelleket. Előbbit legfőképpen az tünteti ki, hogy az önbecsapás jelenségét továbbra is szándékok segítségével igyekszik magyarázni, ellentétben az utóbbival, amely ugyanezt mindenfajta megtévesztési szándékra való hivatkozás nélkül teszi.

Ám a fenti kritériumok nem alkalmazhatóak problémátlanul az önbecsapásra. Ha A és B ugyanazon személy, akkor komoly konceptuális nehézségbe ütközünk. Egyfelől ugyanis ebben az esetben egyetlen személy *egyszerre hiszi azt, hogy p, és ugyanakkor azt is, hogy nem-p*. Nos, *ez egy lehetetlen elmeállapot*. Ez az önbecsapó személyek elmeállapotára vonatkozó rejtély a *stati-*

[1] Vagy legalábbis nem hiszi azt, hogy p.

*kus paradoxon*. Másfelől ha a becsapó és a becsapott ugyanaz a személy, és a megtévesztés per definitionem szándékolt cselekvés, akkor nem világos, miért nem lesz a megtévesztés majdnem biztosan sikertelen folyamat. Ezt az önbecsapás dinamikájára vonatkozó rejtvényt nevezzük *dinamikus paradoxonnak*.[2]

## 2. Intencionális magyarázatok

Az elméletalkotók azon csoportja, amely a személyközi megtévesztés analógiáját (a két kritériummal együtt) meg szeretné tartani, el kell kerülje mind a statikus, mind pedig a dinamikus paradoxont. Az intencionális magyarázati modellek alapvetően két válasszal állnak elő. (1.) Az önmagunk megtévesztésére irányuló szándék csupán *közvetett módon* érvényesül. Ehhez általában *az elme temporális felosztására* szokás hivatkozni, mint azt például Pears (1984) teszi. (2.) Az egész folyamat az önbecsapó személy számára *átláthatatlan módon* zajlik. Ehhez *az elme pszichológiai értelemben vett felosztását* hívják segítségül, így például Donald Davidson (1986, 1997) is. Az első stratégia inkább a dinamikus, a második a statikus paradoxon kiiktatását hivatott elvégezni. Az előbbi esetben a személy az önbecsapó stratégiát (igaz hitestül, szándékostul) egyszerűen elfelejti (mint amikor tíz perccel előrébb állítjuk az óránkat, nehogy elkéssünk), az utóbbi esetben pedig nem világos számára, hogy egymásnak ellentmondó propozíciókat hagy jóvá.

Vegyünk egy példát (Scott-Kakures 2013, a példát módosítottuk) Alfonz, a kiváló, de magányos matematikus megtudja, hogy idős korában minden valószínűség szerint Alzheimer-kórban fog szenvedni. Ha ez a végkimenetel elkerülhetetlen, akkor legalább szeretné időskori napjait abban a hitben tölteni, hogy ő bizony nem is volt magányos. Ezért hát hamis naplóbejegyzéseket, fotókat készít. Ebben az esetben sem a statikus, sem a dinamikus paradoxon nem merül fel. A stratégiáról az idők során megfeledkezik, a két, egymásnak ellentmondó hit pedig egymástól időben olyan távol van, hogy azok inkonzisztenciáját feltárni minden bizonnyal nem áll módjában. E példában a két hit *diakrón* elválasztásával sikerül a statikus paradoxon elkerülése. A *szinkrón* inkonzisztencia azonban szintén nem tűnik lehetetlennek: előfordulhat (főleg bonyolult propozícióhalmazok esetén), hogy nem látjuk be hiteink ellentmondásosságát, mivel nem látjuk át milyen következtetéseket kellene az egyes hitekből levonnunk. Az önbecsapás esetén még az is előfordulhat, hogy valaki azt hiszi, hogy *p* és ugyanakkor azt is, hogy *nem-p*, a kettő konjunkcióját azonban mégsem hiszi (Davidson 1986). „Nem mindig adunk össze kettőt meg kettőt" (Mele 1997, 102. o.).

[2] A két paradoxon részletes tárgyalásához lásd Mele (1998).

Az elme időbeli és/vagy pszichológiai felosztásának elmélete talán képes fogalmi eszközöket biztosítani a fentiekben kitűzött feladatok elvégzéséhez. Mindazonáltal e magyarázatok több, nehezen megválaszolható kérdést szülnek. Például nehéz a szándékot tartalma alapján individuálni. Kialakítani egy, az önbecsapásra irányuló szándékot, igencsak „őrült választásnak" (Lazar 1999) tűnik. Ráadásul mindig fenn fog állni annak a gyanúja, hogy itt teljesen *önkényes, ad hoc szándéktulajdonítás* árán sikerült a paradoxonoktól megszabadulnunk.

## 3. Nem-intencionális magyarázatok

Az intencionális magyarázati modellek fenntartásának nehézségei a filozófusok egy csoportját – köztük Alfred Mele-t is – arra a következtetésre juttatta, hogy az önbecsapást a személyközi megtévesztés mintájára konceptualizáló filozófiai trend egész egyszerűen félrevezető.[3] Mele szerint, még ha a szándékolt önbecsapás nem is tűnik lehetetlennek, nem valószínű, hogy a hétköznapi eseteket ennek analógiájára kell elgondolnunk. Ha sikerül olyan elméletet alkotni, amely képes mindenfajta „mentális egzotikumra" történő hivatkozás nélkül megmagyarázni a jelenséget, úgy (Ockham borotvájának elvére hivatkozva) az intencionális magyarázattal szemben inkább az utóbbit kell előnyben részesítenünk. Feladhatjuk az oly sok fejtörést okozó, fentiekben tárgyalt kritériumokat. Amennyiben elvetjük az egymásnak ellentmondó hitek feltételét, akkor a statikus paradoxon nem merül fel. A szándékoltsági kritérium elvetése pedig a dinamikus paradoxont iktatja ki. Nem létező betegségre pedig felesleges gyógyírt keresni.

Azt nem kell tagadni, hogy az önbecsapás a motivált irracionális hitek egy fajtája. Csak az tűnik valószínűtlennek, hogy az ilyen hiteink egy, az önbecsapásra irányuló szándék eredményeképpen jönnének létre. Szándékolt cselekvéseink halmaza nem fedi teljesen motivált viselkedésmintáink halmazát. Az önbecsapó hitek kétségkívül tévesek. Ugyanakkor nem puszta kognitív hiba eredményeképpen állnak elő; a vágyak, motivációk kauzális szerepet játszanak kialakulásukban. Azonban szándékok nélkül miképpen írható le az a jelenség, hogy vágyaink okságilag hatnak hiteinkre?

Hiteket mindenfajta motivációtól függetlenül is sokszor alakítunk ki torzult módon. A kognitív pszichológiai szakirodalomban az utóbbi évtizedekben került a figyelem középpontjába az ilyen, úgynevezett „hideg" torzító mechanizmusok vizsgálata. Hiteink kialakítása során például hajlamosak vagyunk

[3] E tanulmány keretei között csupán Alfred Mele nem-intencionális elméletét vizsgáljuk. Néhány másik teoretikus –elsősorban az intencionális elméletek kapcsán tárgyalt nehézségek által motiválva – szintén a szándékoltság kizárásával igyekszik a jelenséget magyarázni, pl. Lazar (1999), Barnes (1997), Audi (1976).

aránytalanul nagy súlyt fektetni az élénk, könnyebben hozzáférhető információkra (Nisbett–Ross 1980). Mikor egy hipotézist tesztelünk, gyakrabban keresünk a hipotézist megerősítő példákat: ez a jelenség az úgynevezett *konfirmációs torzítás* (Baron, 1988, Nisbett–Ross 1980). E mechanizmusok se nem szándékoltak, se nem motiváltak; hamis hiteink kialakulásában azonban oksági szerepet játszhatnak. Ezeket azonban vágyaink *felerősíthetik*, az önbecsapó hiteink sokszor ily módon állhatnak elő. Hogy milyen motivált, „forró" torzító mechanizmusokon keresztül történhet meg a hamis hitek (miszerint *p*) kialakulása, lássuk a következő táblázatot (Mele [1997, 2001] alapján):

| | |
|---|---|
| 1. Negatív félreértelmezés | A vágy, hogy *p* igaz legyen, arra készteti *S*-t, hogy a *p* ellen szóló adatokat úgy értelmezze, mint amelyek nem szólnak *p* ellen. Ha a vágy nem játszana közre, az adatokat *p* ellen szólónak tekintené. |
| 2. Pozitív félreértelmezés | A vágy, hogy *p* igaz legyen, arra készteti *S*-t, hogy a *p* ellen szóló adatokat *p* mellett szólóknak tekintse. Ha a vágy nem játszana közre, az adatokat *p* ellen szólónak minősítené. |
| 3. Szelektív figyelem | A vágy, hogy *p* igaz legyen, arra készteti *S*-t, hogy figyelmét (szándékolatlanul vagy szánt szándékkal) a *p* mellett szóló adatokra összpontosítsa, míg a *p* ellen szóló adatokat negligálja (szándékolatlanul vagy szánt szándékkal). |
| 4. Szelektív bizonyítékgyűjtés | A vágynak következtében, hogy *p* igaz legyen, *S* úgy irányítja saját bizonyítékgyűjtő tevékenységét (szándékosan vagy szándékolatlanul), hogy minél többször ütközzön *p*-t igazoló bizonyítékokba, és szisztematikusan elkerülje *nem-p* mellett szóló evidenciákat. |

A bevezetőben tárgyalt példákban közrejátszhatnak ezek a mechanizmusok, s ezek eredményeképp hiheti Sam tévesen azt, hogy Sally szerelmes belé, illetve az anya azt, hogy a gyermeke ártatlan. Sam tekintheti úgy Sally elutasítását, mintha a lány csupán kéretné magát (pozitív és negatív félreértelmezés), vagy eltekinthet a határozott elutasítás tényétől, mint a kérdés szempontjából valójában irreleváns tényezőtől (szelektív figyelem és bizonyítékgyűjtés). Hasonlóképpen: az anya tekintheti úgy a bizonyítékokat, mint annak evidenciáját, hogy fiát csőbe húzták (pozitív és negatív félreértelmezés). Hosszan elidőzhet továbbá azon, hogy (szerinte alaptalanul megvádolt) fia mindig is milyen tisztességes, becsületes ember volt (szelektív figyelem és bizonyítékgyűjtés). Azonban – hangsúlyozza Mele – még ha szándékosan is terelik el figyelmüket a *p* ellen szóló bizonyítékoktól, még ha szándékosan is

fordulnak a *p* mellett szóló bizonyítékok irányába, mindez nem jelenti azt, hogy a viselkedésükből származó adatok indokolttá tennék, hogy önmaguk megtévesztésére irányuló szándékot tulajdonítsunk nekik.

Mindennapi hipotézis-tesztelésünk során – úgy tűnik – nem is elsősorban arra törekszünk, hogy igaz hitekre tegyünk szert, sokkal inkább a költséges hibák elkerülése a cél (Friedrich 1993). Ha ez igaz, akkor jó érvnek tűnik a mellett, hogy az önbecsapás jelenségét a nem-intencionális modell keretei között magyarázzuk. A vágyaink, motivációs állapotaink egész egyszerűen képesek arra, hogy bizonyos hiteknek az úgynevezett *elfogadási küszöbét* (Liberman–Trope 1996) alakítsák (jelenlegi példáink esetén azt csökkentsék).[4] Nincs abban semmi rejtélyes, hogy Sam számára, pillanatnyi motivációs állapota következtében annak a hitnek, hogy Sally szerelmes belé, viszonylag alacsony az elfogadási küszöbe. Az anya esetében szintén nem kell túl nagy empátia ahhoz, hogy belássuk, költséges hiba lenne, ha a fiát tévesen hinné bűnösnek. Ez a motivációs bázis a fia ártatlanságába vetett hit elfogadási küszöbét viszonylag alacsonyra teszi.

Tehát a fentiekben azt láttuk, hogy az úgynevezett „forró" torzító mechanizmusok és mindennapi hipotézis-tesztelésünk kognitív pszichológiai elméletei segítségével az önbecsapás tipikus eseteit mindenfajta önbecsapásra irányuló szándékra történő hivatkozás nélkül is képesek vagyunk magyarázni. Hogy önbecsapásról beszélhessünk, csupán a következő elégséges feltételeknek kell együttesen teljesülniük (Mele 1997, 2001):

1. *S* azon hite, hogy *p*, hamis.
2. *S* (a *p* igazságértéke szempontjából releváns, vagy legalábbis annak látszó) adatokat motiváltan torzítja.
3. A torzult bizonyítékkezelés nem deviáns módon[5] eredményezi *S* azon hitét, hogy *p*.
4. A pillanatnyilag *S* rendelkezésére álló bizonyítékok inkább indokolnák a hitet, hogy *nem-p*, mintsem azt, hogy *p*.

Ha ez a modell helytálló, akkor a személyközi megtévesztés modelljét (és vele mind a statikus, mind a dinamikus paradoxont) egyszer és mindenkorra elvethetjük. De vajon helytálló-e?

---

[4] Más esetekben pedig növeljék – az önbecsapás „fordított" változata esetén, ahol éppen olyasmit hiszünk, amit legkevésbé sem szeretnénk, hogy igaz legyen (pl. beteges féltékenység esetén). Ekkor az összes valós bizonyíték, amely amellett szól, hogy házastársunk hűséges, sem elég ahhoz, hogy ne legyünk meggyőződve ennek ellenkezőjéről.

[5] Mele e ponton egy, leginkább a cselekvéselméletben felmerülő vitás kérdést érint. Ez azonban a jelen tanulmány célkitűzéseinek kontextusában nem releváns nehézség (v.ö. Mele [2013]).

Néhány ellenvetés a nem-intencionális elmélettel szemben is megfogalmazódott. Először: nem világos, hogy Mele azt a jelenséget magyarázza-e, amelyet a hétköznapokban önbecsapásnak szoktunk nevezni. Elmélete inkább az úgynevezett *vágyteljesítő gondolkodás* (wishful thinking) jelenségének magyarázatára tűnik alkalmazhatónak. Ez utóbbi esetben a vágy közvetlenül okozza a hamis hitet, hogy *p*. Míg az intencionális keretben – a szándék segítségével – kvalitatív különbséget tudunk tenni a motivált irracionális hitek különböző fajtái között, addig a nem-intencionális keret erre alkalmatlannak mutatkozik.

Nem feltétlenül sikerül továbbá az *önbecsapás fenomenális karakterét* e magyarázat keretei között megőrizni. Az intencionális elmélet keretei között az egymásnak ellentmondó hitek kritériumának megtartásával sikerült magyarázatot adni az e jelenséghez tipikus esetben kapcsolódó *feszültségre*.[6] Nem világos, hogy Mele modellje ezt a feszültséget miképpen tudja megőrizni.

Ugyanígy magyarázati nehézséget jelent az úgynevezett *szelektivitási probléma*. Ha ugyanis egyedül a vágyak hivatottak magyarázni az önbecsapás jelenségét, akkor az szorul magyarázatra, hogy vajon miért nem csapjuk be önmagunkat lépten-nyomon. Vegyünk egy példát: egy lejtőn az autónkban utazva észrevesszük, hogy nem működik a fék. Ekkor az a vágyunk, hogy a fék mégiscsak működjön, beindíthatná azokat a fentebb vázolt torzító mechanizmusokat, melynek következtében arra a hitre tehetnénk szert, hogy a fékkel bizony semmi baj. Ezzel szemben az út mellé húzódunk, és segítséget hívunk.[7] A vágy önmagában aligha képes arra magyarázatot adni, hogy az imént vázolt esetben miért *nem* indult be a teljesen automatikus, az alany részéről semmilyen erőfeszítést nem igénylő önbecsapó folyamat, míg Sam és az anya esetében miért igen. Az intencionális keretben a szándék ismételten képes számot adni a két helyzet különbségéről.

Ezenfelül mikor azt mondjuk, valaki becsapja önmagát, a jelenséghez, ha homályosan is, de valami erkölcsi szempontból elítélendőt társítunk. Ezt a mind a filozófiatörténetben (ahol az önbecsapás szinte kivétel nélkül morálfilozófiai kontextusban merült fel), mind a mindennapi nyelvhasználatban felbukkanó elemet az intencionális elmélet prima facie képes volt megőrizni. Nem világos azonban, hogy a nem-intencionális elméletek helytállósága esetén nem lesz-e a jelenség morális szempontból a legjobb esetben is irreleváns. A

---

[6] Robert Audi (1976, 1997) elmélete – bár szintén nem tartja meg az egymásnak ellentmondó hitek kritériumát, mégis – képes megmagyarázni ezt a jellegzetességet, hiszen ott a konfliktus (még ha nem is doxasztikus természetű) abból az állapotból fakad, miszerint egyfelől tudom (vagy legalábbis hiszem), hogy *nem-p*, másfelől őszintén vallom, hogy *p*.

[7] A szelektivitási problémát először Talbott (1995) vetette fel. Később pedig – Mele „hideg", valamint „forró" torzító mechanizmusai kapcsán – Bermúdez (1997, 2000).

legmarkánsabban ezt a tételt Neil Levy (2004) fogalmazta meg. Az önbecsapás fentiekben tárgyalt „új fogalma" kapcsán azt állítja, hogy amennyiben egyszer lecseréltük az önbecsapás intencionális magyarázatait a Mele-féle nem-intencionális modellre, úgy „a paradigmatikus esetekben a felelősség tulajdonításának nincs konceptuális tere; azonban szükség sincsen rá. Ennek megfelelően meg kellene tennünk a tradicionális fogalomról való lemondás utolsó lépését, és ki kellene hajítanunk az automatikus felelősség-tulajdonítás elemét. Az önbecsapás a tévedés egy fajtája, és nem szükségképpen kapcsolódik jobban hozzá a vétek gondolata, mint a többi intellektuális hibához" (294. o.). A tanulmány második fele erre, a nem-intencionális elméleteket érintő, az erkölcsi felelősséggel összefüggő ellenvetésre igyekszik választ adni.

## 4. Mechanizmus és erkölcsi felelősség. Két stratégia

Jól látható tehát, hogy a Mele nem-intencionális elmélete az önbecsapást egy többé-kevésbé mechanisztikus folyamatként modellezi. És éppen ez a mechanisztikusság az, ami felveti annak gyanúját, hogy a nem-intencionális elmélet elveszi annak lehetőségét, hogy az önbecsapás jelenségét erkölcsileg értékeljük. Hiszen a természeti folyamatokat, vagy az állatok ösztönös viselkedését éppen azért nem tartjuk erkölcsileg elítélhetőnek vagy dicsérendőnek, mert ezek a jelenségek mintegy automatikusan lezajlanak attól függően, hogy milyen a kiinduló helyzet és az azt érő hatások. Mele elmélete mintha pontosan ezt a fajta automatizmust vázolná fel az önbecsapással kapcsolatban.

Neil Levy, mint láthattuk, egyenesen amellett érvelt, hogy ha Mele-nek igaza van, akkor az önbecsapásért – legalábbis az esetek túlnyomó többségében – nem lehetünk erkölcsileg felelősek. Egyedül azon esetek képezhetnek kivételt a filozófus szerint, amikor a cselekvő belátta, vagy be kellett volna látnia, hogy döntése önmaga megtévesztéséhez fog vezetni. Ám Levy szerint azon a szituációk száma elenyésző, ahol ez a feltétel teljesül. Először is – érvel Levy – (i) szinte lehetetlen előre látni, hogy éppen egy olyan élethelyzetbe fogunk kerülni, ahol a különböző ingerek, valamint az erős vágyaink fel fogják hevíteni torzító mechanizmusainkat. Másodszor, az esetek többségében – miután az előrelátás hiánya miatt belesétáltunk a „csapdába" –, (ii) amennyiben elindult az önbecsapó mechanizmus, lehetetlen észlelni, hogy vágyaink által tényértékelő és szelektáló készségeink akadályoztatva vannak, hiszen éppen ezek a készségek segíthetnének abban, hogy az önbecsapást kiszűrjük. Harmadszor pedig, (iii) utólag sincs igazán esély a korrekcióra, mivel nincs lehetőség egyenként felülvizsgálni a hiteinket, a külvilág pedig egyáltalán nem kényszerít rá arra visszajelzéseivel, hogy az összes önbecsapó hitet detektáljuk.

Alapvetően két út közül választhat az, aki ragaszkodik Mele nem-intencionális elméletéhez, és ugyanakkor szeretné megőrizni a jelenséghez tapadó morális intuíciónk érvényességét. Az első lehetőség abban áll, hogy az illető kimutatja, nemcsak azokért a dolgokért lehetünk felelősek, amelyek valahogyan visszavezethetőek szándékos és tudatos döntéseinkre. E helyütt nem ezt az utat járjuk végig,[8] mivel implauzibilisnek tűnik az az álláspont, amely szerint olyan dolgokért vagy cselekedetekért is elmarasztalhatóak vagy dicsérhetőek lehetünk, amelyek *szükségszerű következményei* voltak születési adottságainknak, vagy külső hatásoknak, esetleg a kettő összjátékának. Ha valaki úgy született, hogy eleve hajlamos volt az agresszióra, és célzottan arra nevelték – minden egyéb releváns külső behatást kizárva –, hogy ilyen és ilyen típusú helyzetekben agresszíven viselkedjen, akkor erkölcsileg nem felelős azért, hogy végül is az x típusú helyzetekben az agresszív viselkedést tartja a lehetős legjobb döntésnek.

A másik opció az volna (és a tanulmány célja ennek az alternatívának a létjogosultságát demonstrálni), hogy igyekszünk megmutatni, az önbecsapást a paradigmatikus esetekben mégiscsak vissza lehet vezetni korábbi döntésekre. Ez a stratégia nincs eleve kudarcra ítélve, amit az is bizonyít, hogy voltaképpen már ki is dolgozták. Ez az ún. *láthatatlan kéz elmélet*, amely azt mondja ki, hogy az önbecsapás tudatos döntések sorozatának szándékolatlan következménye (Galleotti 2012).[9] Bár Galleotti úgy gondolja, hogy a láthatatlan kéz elmélete tulajdonképpen egy harmadikutas megoldás az intencionális és a nem-intencionális elméletek mellett, én úgy vélem, ha van is különbség a nem-intencionális és a láthatatlan kéz elméletek között, az csupán árnyalatnyi. Ugyanis már Mele is azt állítja – mint azt fentebb már láthattuk –, hogy sok esetben tudatos döntések vezetnek az önbecsapáshoz.[10] Egy nem-intencionális elmélet egyáltalán nem attól nem-intencionális, hogy teljességgel kizárná a szándékokat az önbecsapásból, hanem attól, hogy elveti az önbecsapásra irányuló szándékot mint magyarázó eszközt az önbecsapás paradigmatikus eseteivel összefüggésben. Ha így értelmezzük a nem-intencionális elméletek lényegét, akkor látható, hogy a láthatatlan kéz elmélet csupán egy fajtája a nem-intencionális elméleteknek, ami voltaképpen abban különbözik Mele elméletétől, hogy minden egyes esetben megköveteli, hogy az önbecsapás tudatos döntések szándékolatlan eredményei legyenek. Valószínűleg praktikusabb, ha – Mele fogalmi apparátusának megfelelően – a motivált, a tények interpretációját és érzékelését nagymértékben torzító mentális folyamatok

[8] Akadnak bőven tanulmányok, amelyek jó érvekkel szolgálhatnak arra nézvést, hogy miért volna mégis érdemes elfogadni ezt a stratégiát. Adams 2013, Sherr 2013.

[9] Galleotti 2012.

[10] Mele 1997, 98. o.; Mele 2001, 18–21. o.

összességét hívjuk önbecsapásnak, tekintet nélkül arra, hogy szándékos döntések részt vettek-e a folyamatban vagy sem, hiszen ez a gyakorlat igazodik ahhoz a mindennapi nyelvhasználathoz, amelyben a jelenségek egészen széles körét szoktuk „önbecsapás"-ként kategorizálni.

Az önbecsapást így különböző fajtákra lehet osztani annak függvényében, hogy a tudatos döntések milyen helyet foglalnak el bennük. Ezeket a típusokat figyelembe véve próbáljuk bemutatni, hogy az önbecsapás nem-intencionális elméletei, az önbecsapás lehetséges típusainak többségét tekintve, nem zárja ki az erkölcsi felelősség lehetőségét.

## 5. Az önbecsapás típusai.

### 5.1. Súlyos és ártatlan önbecsapás

Az önbecsapásfajták kategorizálásához „sorvezetőként" Levy ellenvetéseit fogjuk felhasználni. Levy a iii)-as ellenvetésében azt állította, hogy az önbecsapás paradigmatikus eseteiben az önbecsapó nincs kellőképpen ösztönözve arra, hogy felülvizsgálja hamis hiteit. Valóban, előfordulnak ilyen helyzetek, amikor a motivált hamis hittel való rendelkezés nem oszt nem szoroz, és a külvilág nem ad visszajelzéseket, mert különösebben nincs következménye a tényeknek ellentmondó hamis hitek. Ezekben az esetekben azért nem felelős az individuum, mert nem vizsgálta felül az önbecsapó hamis hiteket. Ám az ilyen önbecsapó hitek többnyire ártalmatlanok, és a mindennapokban sem szoktuk különösebben pellengérre állítani azt, aki ilyenekre tesz szert. Ha egy nem túl edzettnek tűnő középkorú férfi azt állítja, rossz kondíciója ellenére, hogy ő a mai napig remek hosszútávfutó, azt legrosszabb esetben is csak megmosolyogjuk, de azért nem érezzük felelősnek, hogy ezt a hitet – amellyel életkörülményei folytán nem kellett különösebben foglalkoznia – erkölcsi kötelessége lett volna felülvizsgálnia.

Ezeknél az ártatlan motivált hamis hiteknél sokkal fontosabbak erkölcsileg azok, amelyek valamilyen fontos kérdést érintenek, és káros hatásai vannak. Az ilyen típusú, súlyosabb önbecsapásnál a külvilág folyamatosan küld jelzéseket arra vonatkozóan, hogy az illetőnek felül kellene vizsgálni a hiteit. Ráadásul, ha tudatában vagyunk az adott döntés, és a belőle származó hit fontosságának, ez elvileg arra kellene, hogy ösztökéljen, hogy készen álljunk erre a felülvizsgálatra.[11] Ezért Levy (iii)-as ellenvetésének éppen azokra a helyzetekre vonatkoztatva nincs igazán jelentősége, amelyek miatt egyáltalán felvetődik az önbecsapás és a felelősség kapcsolatát feszegető kérdés. Ám ha valamiért a súlyos önbecsapás mégsem kényszeríti ki a negatív visszajelzése-

[11] Nelkin 2012, 134. o. 36. lj.

ket, s nem igazán belátható az alany számára döntésének és a belőle fakadó hitnek a fontossága – valószínűleg az ilyen esetek viszonylag ritkák –,[12] még mindig könnyen lehet, hogy az illető morálisan felelős, ha nem is a hit fenntartásáért, inkább annak kialakulásáért.

### 5.2. Kvázi-mechanisztikus és mechanisztikus önbecsapás

Levy második ellenvetését, amely szerint a tudattalan torzító mechanizmusok lehetetlenné teszik, hogy észleljük ezek működését, úgy kell megválaszolni, hogy rámutatunk, nincs is szükség a torzító mechanizmusok észlelésére ahhoz, hogy a megismerőnek az önbecsapással kapcsolatban erkölcsi felelősséget tulajdonítsunk.

Mele elméletében az önbecsapás eleve nem feltétlenül egészen mechanisztikus – mint erről már volt szó. Mele-nek több példája is van,[13] amellyel szemlélteti, hogy a szándékos döntések miként tölthetnek be fontos szerepet az önbecsapás folyamatában. Én most Mele-nek egy olyan példáját használnám, kisebb módosításokkal, amelyet eredetileg nem erre a célra használt.[14] Dénes elküldött egy cikket egy tudományos folyóiratban, azonban a névtelen bíráló ezt elutasította. Dénes vágyik arra, hogy a bíráló ítélete tévedésen alapuljon, de emellett azt is szeretné, hogy objektíven értékelje a bíráló észrevételeit. Ahogy olvassa a bírálatot, egyszer csak azon kapja magát, hogy kellemetlenül érzi magát, úgy dönt, hogy most inkább az „fontosabb" részekre ugrik, s lassan meggyőzi magát arról, hogy a bíráló félreértette a cikkét. Pedig, ha nem hagyta volna, hogy a kellemetlen érzés eluralkodjon rajta, s kitartóan és alaposan olvasta volna el a bíráló cikkét, Dénes is belátta volna, hogy a bíráló kritikái nagyon is megalapozottak.

Ugyan Dénes nem amellett döntött, hogy becsapja magát, de döntésének, amit a kellemetlenség érzés elhárítása motivált, kulcsszerepe volt abban, hogy végül arra az önbecsapó hitre jutott, hogy a bíráló félreértette őt. Ráadásul Dénes is tudta, hogy egy részletes bírálat szelektív olvasása az ő helyzetében tudományetikai szempontból nem igazán helytálló. Dénes döntése etikailag önmagában is problematikus volt, és nem volt kikényszerítve a kellemetlenség elkerülését parancsoló ösztön által – hiszen jelen voltak „nemesebb" motivációk is. Ha ezek a feltételek teljesülnek: az önbecsapás felé hajtó vágy nem ellenállhatatlan, a cselekvő valahol érzi (vagy éreznie kellene), hogy döntése etikailag nem egészen „tiszta", és mégis valamiféle (mondvacsinált)

[12] Érdekes probléma, hogy a nagy hatalommal rendelkező emberek könnyebben becsaphatják önmagukat, mivel legtöbbször nem merik felhívni a figyelmüket téves helyzetértékelésükre.

[13] Mele 1997, 98. o.; Mele 2001, 18–21. o.

[14] Mele 1997, 94-95. o., 100. o.

indokok alapján amellett az eljárás mellett dönt, amelyek végül az önbecsapáshoz vezetnek, akkor a cselekvő felelős azért, hogy önbecsapó hite kialakult. Ez attól függetlenül igaz, hogy az önbecsapó hit súlyos vagy éppen ártatlan, hogy jönnek-e a később a vélekedés felülvizsgálatát sürgető jelek, vagy sem. És arra sincs szükség, hogy Dénes átlássa, miféle vágyak hajtják őt jobb meggyőződésével ellentétes irányba. A lényeg, hogy Dénes a kellemetlen érzéseket megtapasztalva inkább a könnyebb utat választotta, pedig tudta, hogy ez a döntése nincs egészen rendben.

De mi a helyzet akkor, ha az önbecsapás teljességgel automatikusan játszódik le? Ha az önbecsapáshoz vezető vágy olyannyira erős, hogy a megismerőben nincsen számottevő ellenállás vele szemben? Térjünk vissza megint Dénes példájához. Tegyük föl, hogy Dénes egyáltalán nem tette belsővé azt az etikai elvet, hogy egy tudományos hozzászólás felett csak akkor ítélkezhetünk, ha azt kellőképpen alaposan megismertük. Így Dénesben nem keletkezik semmiféle belső feszültség, amikor döntenie kell, hogy átugorja-e a számára oly fájdalmas kritikákat, a vágy sürgetésére egyszerűen csak elkezdi szelektíven olvasni a bírálatot. Dénes nem kerül igazi döntési szituációba, már-már reflexszerűen cselekszik, s szó sincs arról, hogy akár másképpen is dönthetett/cselekedhetett volna ebben a szituációban.[15]

Dénes hibáztatható? Bár Dénes ebben a példában még kevésbé látja át, motivációi hogyan torzítják észrevétlenül helyzetértékelését, mégis, a válasz, igen. Ennek az az oka, hogy Dénesben egyrészt túl erősek azok a vágyak, amelyek saját énképének őrzésére irányulnak, másrészt túlságosan gyengék azok, amelyek abba az irányba hatnának, hogy Dénes valamilyen szinten hajlamos legyen arra, hogy a valóságot akkor is meg akarja ismerni, ha az számára kellemetlen. Nehezen elképzelhető, hogy tanuló évei folyamán ne hallott volna arról, hogy kutatóként az igazság elfogulatlan szemlélésére kell törekednie. Ahogy az sem valószínű, hogy ne lett volna alkalma szert tenni valamilyen szinten az olyan erényekre, mint az igazságszeretet, a bátorság vagy éppen az empátia,[16] amelyek a mindennapok folyamán arra sarkalhatják az embert, hogy a kellemetlen igazságokkal is szembenézzen, s ellenálljon a csábító hazugságoknak. Bár nem ismerjük Dénest, de elég hihetetlennek hangzik, hogy pusztán neveltetése, a külső hatások, valamint adottságainak tudható be, hogy ennyire érzéketlen a tudományos normák iránt. Joggal

[15] Itt nagyban támaszkodok Robert Kane elméletére, amely szerint közvetlen és szabad kontrollt akkor vagyunk képesek gyakorolni döntéseink alakulása felett, ha több, egymásnak feszültségben lévő motivációs erő is egyszerre jelen van. Kane 1998, 124–152. o.

[16] Az empátia azáltal, hogy a másokkal való törődésre késztet, hozzásegíthet ahhoz, hogy komolyan vegyük mások véleményét.

gyanakodhatunk arra, hogy Dénes múltbéli döntései állhatnak a háttérben.[17] Ezzel pedig el is érkeztünk oda, hogy válaszoljunk Levy (i)-es ellenvetésére. Még ha a fontos döntési helyzet nem is előre látható, elviekben mégis kötelességünk felkészülni arra, hogy ezekben a helyzetekben elkerüljük önmagunk megtévesztését, mivel – az önbecsapás jelenségétől részben függetlenül – amúgy is kötelességünk szert tenni azokra az erényekre, amelyek megakadályozzák, hogy a jelentős döntési helyzetekben az önbecsapásra hajlamosító vágyaink vegyék át felettünk az uralmat.

Láthatjuk, hogy önmagában az önbecsapás teljes automatikussága sem zárja ki, hogy a cselekvő erkölcsi felelősséggel rendelkezzen, amennyiben a cselekvő a korábbi döntései felelősek azért, hogy a mechanizmus lefolyása ennyire elkerülhetetlenné vált bizonyos szituációkban. Természetesen elvileg elképzelhető, hogy szerencsétlen adottságok és/vagy neveltetés áll a hátterében annak, hogy valakiben az önbecsapási hajlam még a fontos helyzetekben is kontrollálhatatlanná válik. Aligha valószínű, hogy az ilyen teljesen automatikus és teljesen elkerülhetetlen esetek volnának az önbecsapás paradigmatikus esetei. Ám ha az önbecsapás semmiképpen sem vezethető vissza az ágens etikailag problematikus jelenbéli vagy múltbéli döntéseire vagy éppen mulasztásaira, akkor nem tarthatjuk az önbecsapót felelősnek. De mivel a súlyos önbecsapásnál valószínűleg nem írható minden a körülmények és a kezdeti adottságok összjátékának számlájára, ezért, még ha Mele elmélete igaz is, ezekben az esetekben joggal valószínűsíthetjük, hogy az önbecsapó felelős saját maga megtévesztéséért.

## 6. Összefoglalás

Elvitathatatlan, hogy a nem-intencionális megközelítés elegánsan felold néhány olyan paradoxont, amire az önbecsapás intencionális elméletei csak nehezen tudnak válaszolni. Ugyanakkor olyan új problémákat vet föl – mint például a szelektivitás-probléma és az erkölcsi felelősséggel kapcsolatos aggályok –, amelyek a korábbi keretben könnyedén kezelhetőek voltak. Írásunk második felében amellett igyekeztünk érvelni, hogy az erkölcsi felelősség ésszerű tulajdonítását az önbecsapás nem-intencionális elmélete nem teszi lehetetlenné, ezért legalábbis az erkölcsi felelősséggel kapcsolatos aggályok miatt nem érdemes elvetni ezt a teóriát.

[17] Tehát itt az erkölcsi felelősségnek egy lánc-elméletét (trace-theory) alkalmazom. Az erkölcsi felelősség lánc-elméletei azt állítják, hogy sok esetben akkor is felelősek lehetünk, ha a konkrét szituációban nem gyakoroljuk az erkölcsi felelősséghez elégséges kontrollt, amennyiben a kontroll elvesztése, illetve az, hogy kontroll nélkül hogyan cselekszünk az adott szituációban, korábbi döntéseink eredménye, ahol még birtokoltuk a szükséges kontrollt.

BIBLIOGRÁFIA


Audi, R., 1976, „Epistemic Disavowals and Self-Deception," *The Personalist*, 57. 378-385.

Barnes, A., 1997, *Seeing through Self-Deception*, New York, Cambridge University Press.

Baron, J.,1988, *Thinking and deciding*, Cambride University Press.

Bermúdez, J. L., 1997, „Defending Intentionalist Accounts of Self-Deception," *Behavioral and Brain Sciences*, 20: 107–8.

Bermúdez, J. L., 2000, „Self-Deception, Intentions, and Contradictory Beliefs," *Analysis* 60(4): 309–319.

Davidson, D., 1985, „Deception and Division," in *Actions and Events*, E. LePore and B. McLaughlin (szerk.), New York, Basil Blackwell.

Davidson, D., 1997, „Who Fools Who?" In. *Self-Deception and Paradoxes of Rationality*, J.-P. Dupuy (szerk.) Stanford, California.

Friedrich, J., 1993, „Primary error detection and minimization (PEDMIN) strategies in social cognition: a reinterpretation of confirmation bias phenomena," *Psychological Review* 100: 298-319.

Galleotti, A. E., 2012, „Intentional-Plan or Mental Event?" *Humana Mente,* 5. évfolyam 20. szám, 41–66. o.

Huoranszki, F., 2011, *Freedom of the Will: A Conditional Analysis,* New York, Routledge.

Kane, R., 1998, *Significance of Free Will*, Oxford, Oxford University Press.

Kirsch, J., 2005, „What's so Great about Reality?" *Canadian Journal of Philosophy*, 35. évfolyam 3. szám, 407–427.

Lazar, A., 1999, „Deceiving Oneself Or Self-Deceived?" *Mind*, 108: 263–290.

Levy, N., 2004, „Self-Deception and Moral Responsibility," *Ratio (new series)*, 17: 294–311.

Liberman, N. & Trope, Y., 1996, „Social hypothesis-testing: Cognitive and motivational mechanisms," in E. T. Higgins & A. W. Kruglanski (szerk.), *Social psychology: Handbook of basic principle,* New York, Guilford Press.

Mele, A. R., 1997, „Real Self-deception," *Behavioral and Brain Sciences*, 20. évfolyam, 91–136.

Mele, A. R., 1998, „Two Paradoxes of Self-Deception" in Jean-Pierre Dupuy (szerk.), *Self-Deception and Paradoxes of Rationality*, CSLI Publications.

Mele, A. R., 2001, *Self-deception Unmasked*, Princeton–Oxford, Princeton University Press.

Mele, A. R., 2013,"When Are We Self-Decieved?" *Humana Mente,* 5(20): 1-17.

Nelkin, D. K., 2012, *Responsibility and Self-deception: A Framework. Humana Mente,* 5(20): 117–140.

Nisbett, R.–Ross, L., 1980, Human Inference: Strategies And Shortcomings of Social Judgement. Pretence-Hall.

Pears, D., 1984, Motivated Irrationality, New York, Oxford University Press.

Scott-Kakures, D., 2013, „Can You Succeed in Intentionally Decieving Yourself?" *Humana Mente* 5(20): 17-41.

Talbott, W. J., 1995, „Intentional Self-Deception in a Single Coherent Self," *Philosophy and Phenomenological Research*, 55: 27–74.